Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks.

By: Paul J. Silvia

Paul J. Silvia (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity, 6(1),* 24-30.

Abstract:

The present research examined the reliability of three types of divergent thinking tasks (unusual uses, instances, consequences/implications) and two types of subjective scoring (an average across all responses vs. the responses people chose as their top-two responses) within a latent variable framework, using the maximal-reliability H statistic. Overall, the unusual uses tasks performed the best for both scoring types, the instances tasks had less reliable scores, and the consequences tasks had poor reliability and convergence problems. The discussion considers implications for test users, differences between average scoring and top-two scoring, and the problem of whether divergent thinking tasks are interchangeable.

Keywords: creativity | divergent thinking | measurement | reliability | latent variable models | psychology

Article:

Divergent thinking tasks, for better or worse, are among the most widely used tools to measure creativity (Kaufman, Plucker, & Baer, 2008), particularly the ability to generate creative ideas (Silvia et al., 2008). As open-ended tasks, they present researchers with an almost paralyzing number of possibilities for scoring, and the decades of research on divergent thinking have provided many ways of creating tidy numerical scores from people's words and scribbles (see Michael and Wright, 1989, Plucker and Renzulli, 1999 and Silvia et al., 2008). For verbal tasks, these methods split into two kinds: objective scoring methods (e.g., Wallach & Kogan, 1965) and subjective scoring methods (e.g., Silvia et al., 2008 and Silvia et al., 2009a). The objective methods have been extensively studied and form the basis of the best-known assessment approaches, such as Wallach and Kogan's (1965) tasks and the Torrance Tests of Creative Thinking (Torrance, 2008). Subjective methods, in contrast, have received less attention, probably because they take more research resources (Silvia, Martin, et al., 2009) and because

researchers often confuse “objective” with “valid” (Webb, Campbell, Schwartz, & Sechrest, 1966).

To expand the growing literature on subjective scoring, the present research examines the score reliability of subjective scoring approaches. Reliability is fundamental to the enterprise of research, and it is essential to understand the reliability of scores generated by new methods. As we will see, subjective scoring methods require different approaches to estimating reliability, particularly methods that separate variance due to traits and to raters. After considering the estimation of reliability within latent variable models, we estimate the reliability of two subjective scoring approaches for three kinds of divergent thinking tasks.

1. The reliability of subjective ratings

Using subjective ratings complicates the ordinarily simple issue of estimating reliability. Because more than one rater ought to be used (Silvia et al., 2008), using raters adds a facet to the model: the researcher will have several divergent thinking tasks, each scored by several raters. The problem is that the raters’ scores are not independent: scores from one rater are more likely to be more similar to scores from that rater than from other raters, so ignoring the rater facet (such as by treating each rater as an independent item in a Cronbach's alpha analysis) will yield misleading estimates of reliability.

The dependence of the raters’ scores can be modeled with latent variable methods, such as the multitrait–multimethod confirmatory factor analysis (CFA) model shown in Fig. 1. Each divergent thinking task is represented by a latent variable with raters’ scores as the indicators, and the dependence of the raters’ scores is represented with covariances between the residual variances. This CFA model nicely depicts the study's measurement structure, but it takes the estimation of reliability outside of the realm of Cronbach's alpha computed on observed scores. Furthermore, researchers may wish to specify measurement models for which traditional reliability estimates cannot be applied. For example, if researchers have many tasks, they can estimate a higher-order latent variable that represents the construct that explains why the tasks covary, such as a higher-order Creativity variable (e.g., Silvia, Nusbaum, Berg, Martin, & O’Connor, 2009). One would like to know the reliability of the higher-order construct, but its indicators are themselves latent variables, not observed scores.

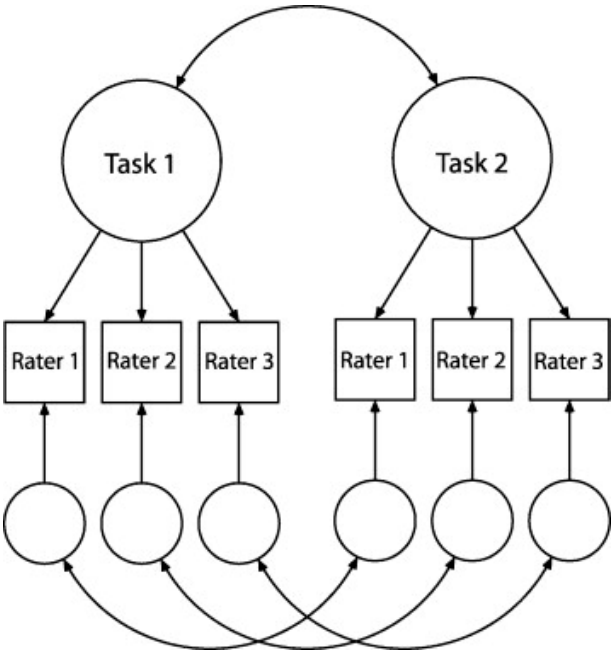

Fig. 1. Depiction of a CFA model that includes several divergent thinking tasks and several raters.

When using subjective ratings, researchers thus need an estimate of reliability that is suitable for latent variable models. Latent variable modeling is a sophisticated way of modeling error, such as error variance due to raters, but it does not absolve researchers of the need to consider the reliability of assessment. One might argue that this need is greater in creativity research, which faces the task of measuring a truly complicated construct.

For latent variables, reliability can be estimated with coefficient *H*, known as *construct reliability* (Hancock & Mueller, 2001) and *maximal reliability* (Drewes, 2000 and Raykov, 2004). *H* can be understood, according toHancock and Mueller (2001), as the "proportion of variability in the construct explainable by its own indicator variables" (pp. 202–203). If the latent variable were regressed on its predictors, *H* would represent the model'sR^2, the percentage of variance that the predictors (the indicators) explain in the outcome (the latent variable).

The value of *H* for a latent variable with *k* indicators and standardized loadings *a* is defined by

$$H = \frac{\sum a^2/(1-a^2)}{1+\sum a^2/(1-a^2)}$$

For example, a latent variable with four indicators with standardized loadings of .90, .85, .80, and .75 yields an *H* = .91.
Specifically,*H* = (.81/.19) + (.72/.28) + (.64/.36) + (.56/.44)/1 + [(.81/.19) + (.72/.28) + (.64/.36) + (.56/.44)] = .91.

Coefficient *H* has some useful properties for estimating reliability for latent variables. First, the sign of the indicator loading is immaterial to *H* because each indicator's standardized weight is squared. Some reliability coefficients require that all indicators have the same sign (e.g., Cronbach's alpha), but the direction of a relationship in psychology is often due to arbitrary scoring decisions. In a CFA model of creativity, for example, researchers would expect an indicator that captures "conventionality" to have a negative loading.

Second, *H* will never be lower than the reliability of the most reliable indicator. As a result, adding indicators will always increase *H* provided that they have non-zero loadings. This contrasts with many measures of reliability, such as Cronbach's alpha, which decline when weaker indicators are added. In the earlier example, for instance, the indicator with the loading of .90 has a reliability of $.90^2$, or .81, which is lower than the *H* of .91. This property of *H* stems from the assumption in latent variable modeling that adding indicators always adds additional information about the construct (see Hancock & Mueller, 2001). Stated differently, a latent variable cannot convey less information than its most informative indicator. If it did, then it would paradoxically be better to use the observed indicator instead of the latent variable.

Because a model's H would not be lower than its best indicator, H will generally be higher than Cronbach's alpha. To understand the relationship between alpha and H, we can develop an empirical example. Consider a study that administered six divergent thinking tasks and used fluency (the number of responses) as the measure of creativity, which is common in divergent thinking research. The present study, not by coincidence, administered six tasks, and Cronbach's alpha for the fluency scores (estimated via SPSS with ordinary least-squares) was .802. To obtain alpha via a CFA, we specify the six scores as indicators of a single latent variable, and we specify that the standardized loadings are tau-equivalent (i.e., constrained to be equal). For this CFA, the H value (estimated via Mplus with maximum-likelihood) was H = .802, an identical estimate. To estimate maximal reliability, we simply drop the tau-equivalence constraint, which allows each indicator to have a unique loading. For this CFA, the H value was .851, a higher value. H thus contains alpha as a special case.

Like alpha, then, H is a function of the number of indicators and the size of their loading; unlike alpha, H does not constrain the loadings to be tau-equivalent, and it will increase as indicators are added. In applied research, similar standards and cut-offs are used to judge the acceptability of an H value, such as a level of .70 or .80.

2. The present research

The versatility of H makes it ideal for creativity research, which will often need complex CFA models to represent designs with several facets, such as task types, raters, and time points. In the

present research, I considered the reliability of two types of subjective scoring for three types of tasks. The first scoring method, known simply as average scoring, asks raters to score each response to a task and then averages the scores—all responses thus contribute to the score. The second scoring method, known as top-two scoring, asks participants to pick their two responses that they think are their most creative. The raters' scores of these top-two responses are averaged for an overall score (Silvia et al., 2008). Both scoring methods have fared well in research, but the top-two approach generally has larger effect sizes, probably because it omits the mundane responses that everyone generates (Silvia et al., 2008 and Silvia et al., 2009b).

Both scoring methods were applied to three types of tasks: unusual uses tasks (generating creative uses for common objects), instances tasks (generating creative examples from a common category), and consequences tasks (generating creative implications of a hypothetical event). It is important to consider several task types because it is currently controversial whether different kinds of tasks are interchangeable (Almeida et al., 2008, Clapham, 2004 and Kuhn and Holling, 2009). Stated differently, divergent thinking tasks may be fixed rather than random, so researchers may not be willing to generalize across tasks. It seems likely that some kinds of divergent thinking tasks recruit different skills, strategies, or knowledge, so it is important to understand the reliability of assessment for several kinds of tasks.

3. Method

3.1. Participants

The sample consisted of 226 undergraduates (178 women, 48 men) who took part in the Creativity and Cognition Project, a study of creativity, personality, and intelligence. Most of the students were 18 or 19 years old (82%), and less than 3% had declared psychology as a major. Nursing (31%) was the most prevalent major, followed by undecided (11%) and biology (5%).

3.2. Method

People completed six divergent thinking tasks: two Unusual Uses tasks (unusual uses for a brick and for a knife), two Instances tasks (instances of things that are round and that make a noise), and two Consequences tasks (consequences of not needing to sleep and of everyone shrinking to 12 in. in height). The tasks were completed in the following order: brick, round, sleep, knife, noise, and inches. For each task, people were told that it concerned creativity and that they ought to try to be creative. Not all divergent thinking assessment uses "be creative" instructions, but research shows that instructing respondents to be creative increases the validity of divergent thinking scores (e.g., Harrington, 1975 and Katz and Poag, 1979).

People had three minutes to complete each task. After the task, they were asked to read their responses and then circle the two that they thought were their best. By asking people to choose their two most creative responses, we can use a kind of maximal assessment: people are asked to be creative and then judged based on the creativity of their best responses (cf. Runco, 1986).

Three undergraduates independently rated each response. The responses were transcribed into a spreadsheet and then sorted alphabetically within task prior to scoring, so the raters were unaware of the person's other responses, whether a response was chosen as a top-two response, idiosyncratic information available from the hard copies (e.g., handwriting), and the scores given by the other raters. Each response was rated on a 1–5 scale (1 = not at all creative, 5 = very creative).

We examined two scoring methods. For average scoring, each response receives a rating, and the average across a rater's score for the task serves as the final score. Average scoring thus uses all of the responses. For top-two scoring, only the rater's scores for the two responses chosen as the best are included; the average across a rater's scores for the top-two responses serves as the final score. As a result, top-two scoring is based on far fewer responses—across all 6 tasks, approximately 25% of the 10,749 usable responses were chosen as top-two responses.1

4. Results

Table 1 displays the descriptive statistics for the six tasks for each scoring method. All models were estimated with Mplus 6.0, using full-information maximum-likelihood with robust standard errors. For each type of task, I specified the confirmatory factor model depicted in Fig. 1. This kind of multitrait–multimethod model—known as a correlated trait, correlated uniqueness model (Brown, 2006 and Lance et al., 2002)—represents rater-specific variability by allowing each rater's residual variances to covary. Ratings are thus a function of people's task score (the latent task variable) and rater-specific features (the residual covariance). Six CFA models were estimated: one for each task type (unusual uses, instances, consequences) and for each scoring type (average scoring, top-two scoring).

Table 1. Descriptive statistics for the six divergent thinking tasks.

	M	*Mdn*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. Brick: average	1.68	1.67	.30	1											
2. Knife: average	1.76	1.72	.34	.449	1										

	M	*Mdn*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
3. Round: average	1.43	1.33	.31	.194	.170	1									
4. Noise: average	1.35	1.31	.20	.137	.208	.178	1								
5. No sleep: average	1.52	1.50	.27	.127	.145	.109	.111	1							
6. 12 in. tall: average	1.39	1.38	.23	.152	.292	.091	.234	.269	1						
7. Brick: top two	1.89	1.83	.47	.745	.332	.117	.093	.137	.139	1					
8. Knife: top two	1.94	1.83	.47	.332	.782	.176	.167	.028	.199	.228	1				
9. Round: top two	1.57	1.50	.46	.178	.099	.731	.098	.151	.080	.194	.148	1			
10. Noise: top two	1.45	1.33	.33	.134	.109	.174	.678	.202	.245	.118	.138	.152	1		
11. No sleep: top two	1.58	1.50	.40	.114	.110	.141	.131	.721	.255	.106	.059	.117	.180	1	
12. 12 in. tall: top two	1.47	1.33	.36	.080	.176	.005	.222	.235	.741	.020	.146	.003	.235	.189	1

Note: Scores are averaged across the three raters. The raw data and complete correlation matrix are available from the author.

4.1. Unusual uses tasks

The CFA models for the unusual uses tasks fit well−Table 2 displays the fit statistics. For average scoring, the standardized loadings for the three raters were β = .933, .807, and .423 for the Brick task and β = .863, .868, and .473 for the Knife task, yielding *H* values of .90 and .86 (see Table 3). The residual correlations for the raters were small− βs for the three raters were −.285, .08, and .084−so rater-specific influences on the scores were modest.

Table 2. Model fit for the confirmatory factor analyses.

	χ^2	CFI	RMSEA	SRMR
Unusual uses: average scoring	6.75 (p = .24)	.996	.039	.023
Unusual uses: top-two scoring	8.78 (p = .12)	.985	.058	.030
Instances: average scoring	6.56 (p = .25)	.981	.037	.030
Instances: top-two scoring	4.59 (p = .47)	.999	.001	.026
Consequences: average scoring	10.17 (p = .25)	.985	.035	.034

	χ^2	CFI	RMSEA	SRMR
Consequences: top-two scoring	–	–	–	–

Note: Empty cells indicate that the model did not converge to an admissible solution.

Table 3. Construct reliability for the divergent thinking tasks.

	Average scoring	Top-two scoring
Unusual uses: brick	.90	.75
Unusual uses: knife	.86	.79
Instances: round	.64	.63
Instances: noise	.74	.59
Consequences: no sleep	.71	.66
Consequences: 12 in. tall	.87	–

Note: Values are *H* scores. Empty cells indicate that the model did not converge to an admissible solution.

For top-two scoring, the loadings for the three raters were β = .763, .755, and .475 for the Brick task and β = .719, .833, and .492 for the Knife task, yielding *H* values of .75 and .79 (see Table 3). The residual correlations for the raters were near zero−βs for the three raters were −.007, .019, and −.018−so rater-specific influences on the scores were tiny.

4.2. Instances tasks

The CFA models for the instances tasks fit well−Table 2 displays the fit statistics. For average scoring, the loadings for the three raters were β = .636, .629, and .554 for the Round task and β = .835, .359, and .520 for the Noise task, yielding *H* values of .64 and .74 (see Table 3). The residual correlations for the raters were β = .035, .218, and .192, so rater-specific influences on the scores were again modest.

For top-two scoring, the loadings for the three raters were β = .672, .627, and .427 for the Round task and β = .723, .190, and .468 for the Noise task, yielding *H* values of .63 and .59 (see Table 3). The residual correlations for the raters were β = −.224, .164, and .142, so rater-specific influences on the scores were modest.

4.3. Consequences tasks

Unlike the unusual uses and instances tasks, the consequences tasks posed problems with model convergence to an admissible solution. For average scoring, the CFA converged when the three correlations between residual variances were omitted. The model fit of this CFA is shown in Table 2. The loadings for the three raters were β = .697, .768, and .293 for the Sleep factor and β = .562, .931, and .172 for the Inches factor, yielding *H* values of .71 and .87 (see Table 3).

5. General discussion

5.1. Reliability across tasks and scoring methods

Reliability is fundamental to assessment, so it is important to understand the features of the scores developed by new assessment methods. The present research examined the reliability of scores from two scoring methods for three kinds of divergent thinking tasks. Using H, an estimate of construct quality that is suitable for complex latent variable models (Hancock and Mueller, 2001 and Raykov, 2004), the analyses revealed findings with clear implications for researchers interested in subjective scoring methods.

First, regarding scoring methods, average scoring was generally more reliable than top-two scoring (see Table 2). This is not surprising, given that average scores are based on all of the responses, whereas top-two scores for a task are based on only two responses. Average scores would thus be expected to be more stable. A similar finding appeared in a small-sample generalizability analysis (Silvia et al., 2008, Study 1): average scoring yielded scores that were somewhat more reliable than top-two scoring. Because both methods use the same data—the top-two scores are a subset of the average scores—it would be natural for researchers to use and report both methods.

Second, regarding tasks, reliability levels differed substantially across the tasks. On the one end, the unusual uses tasks performed the best overall. For both scoring methods, the unusual uses tasks yielded H values that are suitable for basic research. The instances tasks were intermediate: the models converged, but the H levels were notably lower than the unusual uses tasks. These tasks are perhaps not so poor to recommend against their use, but more work is needed before they could be recommended with confidence. One the other end, the consequences tasks performed poorly. For both scoring methods, the multitrait–multimethod CFA shown in Fig. 1 would not converge to an admissible solution. A tweaked model worked for average scoring, but no suitable model would converge for top-two scoring.

Discussions with the raters for this data set and an earlier data set (Silvia et al., 2008, Study 1) revealed that the consequences task was the hardest to score. Participants commonly gave contradictory responses—for the Sleep task, for example, many people wrote “It would take longer to finish college,” many more people wrote “People could finish college faster,” and some people bizarrely wrote both responses—and the overall creativity ratings were near the floor of the scale. Raters also noted that it was hard to appraise the cleverness and unusualness of responses to events that could never happen. Given the present findings, subjective scoring methods for consequences tasks cannot be recommended.

It is worth emphasizing that these results should not be taken out of their assessment context, such as the instructions and timing parameters for the tasks. For example, we asked participants to try to come up with creative responses, and most divergent thinking research does not do this. Likewise, research is inconsistent in the use of timed versus untimed tasks. Different studies may use the same tasks, but orienting participants to creativity and asking them to pick their most creative responses makes the assessment approach different. To evaluate how such factors affect reliability, researchers with many datasets could combine them as part of a reliability generalization study, which examines how score reliability differs across samples, contexts, and task features (Vacha-Haase, 1998). Creativity research has been criticized over the years for the quality of its assessment, and employing sophisticated approaches to reliability estimation would help address such criticisms.

Similarly, we should note that average scoring and top-two scoring have some important differences. They stem from the same assessment method, but they probably capture different sides of divergent thinking. In particular, top-two scoring asks for people's self-assessment of their most creative responses. As a result, both generative and evaluative aspects of creative thought influence people's scores. Not surprisingly, research shows that people's top-two judgments covary strongly with the raters’ judgments: people seem capable of discerning their better ideas from their worse ideas (Silvia, 2008b). But some people's top-two judgments covaried more strongly than others’ judgments did, consistent with the notion that creative evaluation is a skill that varies between people (Grohman et al., 2006 and Sternberg, 2006). Interestingly, we have consistently found that top-two scoring yields higher effect sizes than average scoring (Silvia et al., 2008, Study 2; Silvia, Nusbaum, et al., 2009) and snapshot scoring, a simple holistic scoring method (Silvia, Martin, et al., 2009).

Broadly speaking, the aim of this program of work is to explore new methods for scoring divergent thinking, not to argue for a single method. Top-two scoring in particular is controversial (see Baer, 2008, Kim, 2008 and Runco, 2008), but the validity of psychological

assessment is largely an empirical question. Divergent thinking research primarily uses scoring methods proposed in the 1960s, so creativity researchers, if anyone, ought to find value in exploring novel scoring methods and statistical models.

5.2. Are types of divergent thinking tasks interchangeable?

Divergent thinking research commonly administers many kinds of tasks and then combines the scores, such as by averaging or summing. The tasks can be diverse: many kinds of verbal tasks have been used, and the verbal and figural classes themselves capture different facets of creativity. An issue, however, is whether researchers should be willing to view divergent thinking tasks as interchangeable and equivalent. Scores from one kind of task may not be equivalent from scores from other kinds: the tasks might capture different facets of creativity or different knowledge, strategies, and abilities needed for successful task performance.

Recent work has highlighted differences between tasks. In a small-sample generalizability analysis of three tasks (one unusual uses, one instances, and one consequences task), Silvia et al. (2008, Study 1) found that the tasks appeared to be fixed rather than random, which indicates that their scores are not interchangeable. In an analysis of several data sets, Almeida et al. (2008) found many differences according to task and method, which again indicates a lack of generality. Finally, several studies have found low correlations and salient differences between the verbal and figural classes of divergent thinking tasks (e.g., Clapham, 2004 and Kuhn and Holling, 2009).

It is important to keep in mind that task-specific differences do not necessarily mean that one or more kinds of tasks yield invalid scores. Instead, the different task types could be capturing different facets and dynamics of creative ideation. Just like a broad construct like intelligence, a broad construct like creativity has several layers or facets. At the same time, the task-specific features indicate that researchers should be cautious in combining scores from different task types and that future work should try to understand the psychological basis of these psychometric differences.

The question, then, is was might the psychological basis be? To date, creativity work has not generally examined the kinds of abilities and strategies that undergird divergent thinking tasks. They have been treated as tests—assessment tools that yield scores—rather than tasks—ways of revealing cognitive processes. One could only speculate, but I suspect that executive processes

are relatively more important in unusual uses tasks, which require people to inhibit accessible, obvious uses that interfere with generating novel uses (Gilhooly, Fioratou, Anthony, & Wynn, 2007). For instances tasks, in contrast, the kinds of strategies that would create high scores probably resemble strategies used in verbal fluency tasks, particularly semantic fluency tasks that ask people to enumerate instances of a semantic category (Troyer & Moscovitch, 2006). And for consequences tasks, it is hard to say what kinds of strategies would be successful, given that in the present work they performed poorly psychometrically.

In short, it seems possible that unusual uses tasks are relatively more executive, whereas instances tasks are relatively more associationistic. We would thus expect stronger contributions of executive abilities (e.g., fluid intelligence and working memory capacity) to unusual uses scores than instances scores, and we would expect different kinds of strategies to be effective. The different cognitive bases of the tasks would be one reason why they seem not to generate interchangeable scores, but such predictions naturally await future research.

Acknowledgements

I thank Chris Barona, Josh Cram, Karl Hess, Jenna Martinez, and Crystal Richard for their help in collecting the original data.

References

L.S. Almeida, L.P. Prieto, M. Ferrando, E. Oliveira, C. Ferrándiz. Torrance Test of Creative Thinking: The question of its construct validity. Thinking Skills and Creativity, 3 (2008), pp. 53–58

J. Baer. Commentary: Divergent thinking tests have problems, but this is not the solution. Psychology of Aesthetics, Creativity, and the Arts, 2 (2008), pp. 89–92

T.A. Brown. Confirmatory factor analysis for applied research. Guilford, New York (2006)

M.M. Clapham. The convergent validity of the Torrance Tests of Creative Thinking and creative interest inventories. Educational and Psychological Measurement, 64 (2004), pp. 828–841

D.W. Drewes. Beyond the Spearman-Brown: A structural approach to maximal reliability. Psychological Methods, 5 (2000), pp. 214–227

K.J. Gilhooly, E. Fioratou, S.H. Anthony, V. Wynn. Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. British Journal of Psychology, 98 (2007), pp. 611–625

M. Grohman, Z. Wodniecka, M. Kłusak. Divergent thinking and evaluation skills: Do they always go together? Journal of Creative Behavior, 40 (2006), pp. 125–145

G.R. Hancock, R.O. Mueller. Rethinking construct reliability within latent variable systems. R. Cudeck, S. du Toit, D. Sörbom (Eds.), Structural equation modeling: Present and future, Scientific Software International, Lincolnwood, IL (2001), pp. 195–216

D.M. Harrington. Effects of explicit instructions to “be creative” on the psychological meaning of divergent thinking test scores. Journal of Personality, 43 (1975), pp. 434–454

A.N. Katz, J.R. Poag. Sex differences in instructions to “be creative” on divergent and nondivergent test scores. Journal of Personality, 47 (1979), pp. 518–530

J.C. Kaufman, J.A. Plucker, J. Baer. Essentials of creativity assessment. Wiley, Hoboken, NJ (2008)

K.H. Kim. Commentary: The Torrance Tests of Creative Thinking have already overcome many of the perceived weaknesses that Silvia et al.’s methods are intended to correct. Psychology of Aesthetics, Creativity, and the Arts, 2 (2008), pp. 97–99

J.-T. Kuhn, H. Holling. Measurement invariance of divergent thinking across gender, age, and school forms. European Journal of Psychological Assessment, 25 (2009), pp. 1–7

C.E. Lance, C.L. Noble, S.E. Scullen. A critique of the correlated trait–correlated method and correlated uniqueness models for multitrait–multimethod data. Psychological Methods, 7 (2002), pp. 228–244

W.B. Michael, C.R. Wright. Psychometric issues in the assessment of creativity. J.A. Glover, R.R. Ronning, C.R. Reynolds (Eds.), Handbook of creativity, Plenum, New York (1989), pp. 33–52

Psychometric approaches to the study of human creativity. R.J. Sternberg (Ed.), Handbook of creativity, Cambridge University Press, New York (1999), pp. 35–61

T. Raykov. Estimation of maximal reliability: A note on a covariance structure modelling approach. British Journal of Mathematical and Statistical Psychology, 57 (2004), pp. 21–27

M.A. Runco. Maximal performance on divergent thinking tests by gifted, talented, and nongifted students. Psychology in the Schools, 23 (1986), pp. 308–315

M.A. Runco. Commentary: Divergent thinking is not synonymous with creativity. Psychology of Aesthetics, Creativity, and the Arts, 2 (2008), pp. 93–96

P.J. Silvia. Another look at creativity and intelligence: Exploring higher-order models and probable confounds. Personality and Individual Differences, 44 (2008), pp. 1012–1021

P.J. Silvia. Discernment and creativity: How well can people identify their most creative ideas? Psychology of Aesthetics, Creativity, and the Arts, 2 (2008), pp. 139–146

P.J. Silvia, C. Martin, E.C. Nusbaum. A snapshot of creativity: Evaluating a quick and simple method for assessing divergent thinking. Thinking Skills and Creativity, 4 (2009), pp. 79–85

P.J. Silvia, E.C. Nusbaum, C. Berg, C. Martin, A. O'Connor. Openness to experience, plasticity, and creativity: Exploring lower-order, higher-order, and interactive effects. Journal of Research in Personality, 43 (2009), pp. 1087–1090

P.J Silvia, B.P. Winterstein, J.T. Willse, C.M. Barona, J.T. Cram, K.I. Hess et al. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. Psychology of Aesthetics, Creativity, and the Arts, 2 (2008), pp. 68–85

R.J. Sternberg. The nature of creativity. Creativity Research Journal, 18 (2006), pp. 87–98

E.P. Torrance. Torrance Tests of Creative Thinking: Norms-technical manual, verbal forms A and B. Scholastic Testing Service, Bensenville, IL (2008)

A.K. Troyer, M. Moscovitch. Cognitive processes of verbal fluency tasks. A.M. Poreh (Ed.), The quantified process approach to neuropsychological assessment, Taylor & Francis, Philadelphia (2006), pp. 143–160

T. Vacha-Haase. Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58 (1998), pp. 6–20

M.A. Wallach, N. Kogan. Modes of thinking in young children: A study of the creativity–intelligence distinction. Holt, Rinehart, & Winston, New York (1965)

E.J. Webb, D.T. Campbell, R.D. Schwartz, L. Sechrest. Unobtrusive measures: Nonreactive research in the social sciences. Rand McNally, Oxford, UK (1966)

1 Some of the divergent thinking data from this large data set have been analyzed in prior publications. The Brick and Knife data have been analyzed in studies of how personality (Silvia et al., 2008, Study 2) and intelligence (Silvia, 2008a) predict divergent thinking and in a study of how top-two Brick and Knife scores fare against a simpler scoring method (Silvia, Martin, et al., 2009). The two Instances tasks were analyzed as part of a within-person analysis of participant–rater concordance (Silvia, 2008b), and the two Consequences tasks have not been published. Individual differences in the Instances and Consequences tasks have thus not been published to date.